First Year Examination
Department of Statistics, University of Florida
May 14, 2004, 8:00 am - 12:00 noon

Instructions:

1. You have four hours to answer questions in this examination.
2. You must show your work to receive credit.
3. There are 10 problems of which you must answer 8.
4. Only your first 8 problems will be graded.
5. While the 10 questions are equally weighted, some problems are more difficult than others.
6. The parts within a given question are not necessarily equally weighted.
7. You are allowed to use a calculator.
8. **Write only on one side of the paper, and start each question on a new page.**

The following abbreviations are used throughout:

- ANOVA = analysis of variance
- cdf = cumulative distribution function
- iid = independent and identically distributed
- mgf = moment generating function
- MSE = mean squared error
- ML = maximum likelihood
- MP = most powerful
- pdf = probability density function
- pmf = probability mass function

You may use the following facts/formulas without proof:

Order Statistics: If $X_1, \ldots, X_n$ are iid with common pdf $f(x)$ and common cdf $F(x)$, then the joint density of $X_{(1)}$ and $X_{(n)}$ is

$$f_{1,n}(u,v) = \begin{cases} n(n-1)f(u)f(v)\,[F(v)-F(u)]^{n-2} & -\infty < u < v < \infty \\ 0 & \text{otherwise} \end{cases}$$

Linear Combinations of Independent Normals: Let $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$. If $a_1, \ldots, a_n$ are constants, then the random variable $\sum_{i=1}^n a_i X_i$ has a normal distribution.

1. Suppose that the random variables $Y_1, \ldots, Y_n$ satisfy

$$Y_i = x_i\beta + \varepsilon_i \ , \quad i = 1, \ldots, n$$

where $x_1, \ldots, x_n$ are fixed constants, $\varepsilon_1, \ldots, \varepsilon_n$ are iid $\mathrm{N}(0, \sigma^2)$ and σ^2 is known.

(a) Find the ML estimator of β, call it $\hat{\beta} = \hat{\beta}(Y)$.

(b) Find the distribution of $\hat{\beta}$.

(c) Find the distribution of the alternative estimator of β given by

$$\tilde{\beta} = \tilde{\beta}(Y) = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i} \ .$$

(d) Find the posterior distribution of β under a normal prior with mean 0 and variance $\tau^2/\big(\sum_{i=1}^n x_i^2\big)$.

(e) Show that the posterior expectation of β, call it $\beta_B = \beta_B(Y)$, can be written as a simple function of $\hat{\beta}$.

(f) Compare these three estimators using MSE. Does any one of the three dominate the others? Can any one of the three be ruled out?

2. Let X be a continuous random variable with pdf $f(x)$ and cdf $F(x)$. Assume that $\mathrm{E}X < \infty$.

(a) Define the *sparsity function*, $s : (0, 1) \to \mathbb{R}$, to be the derivative of the quantile function; that is,

$$s(\alpha) = \frac{d}{d\alpha} F^{-1}(\alpha) \ .$$

Using only the chain rule, show that

$$s(\alpha) = \frac{1}{f\big(F^{-1}(\alpha)\big)} \ .$$

(Hint: Think about the Inverse Function Theorem.)

(b) Find the sparsity function when $X \sim \text{Uniform}(0, 1)$ and when $X \sim \text{Exp}(1)$. Why do you think the term "sparsity" is appropriate?

(c) Fix $\alpha \in (0, 1)$ and let X_α be the random variable with cdf given by $F_\alpha(t) = P\big(X \leq t | X > F^{-1}(\alpha)\big)$. Find a closed-form expression for the pdf of X_α.

(d) Find the pdf of $Y_\alpha = X_\alpha - F^{-1}(\alpha)$.

(e) What is the distribution of Y_α when $X \sim \text{Exp}(1)$? Is the answer surprising? Why?

3. This problem concerns consistent estimation of the parameters of a normal distribution.

(a) Derive Chebychev's inequality; that is, show that if X is a random variable, $g(\cdot)$ is a nonnegative function, and $r > 0$, then

$$\Pr(g(X) \geq r) \leq \frac{\mathrm{E}g(X)}{r} \; .$$

Let $Y_1, Y_2, Y_3, \dots$ be an iid sequence of random variables from a distribution with a finite second moment. Let $\mathrm{E}Y_1 = \mu$ and $\mathrm{Var}\, Y_1 = \sigma^2$. Define $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$.

(b) (Weak Law of Large Numbers.) Show that $\overline{Y}_n \xrightarrow{P} \mu$.

(c) Show that if $\mathrm{Var} S_n^2 \to 0$ as $n \to \infty$, then $S_n^2 \xrightarrow{P} \sigma^2$.

(d) Suppose that $Y_1, Y_2, Y_3, ...$ are iid $\mathrm{N}(\mu, \sigma^2)$. Does $\overline{Y}_n \xrightarrow{P} \mu$? And does $S_n^2 \xrightarrow{P} \sigma^2$?

4. Suppose that $X_1, \dots, X_n$ are iid random variables such that

$$P(X_1 = x) = p(1-p)^x \;\text{ for }\; x = 0, 1, 2, \dots$$

where $p \in (0, 1)$.

(a) Does the the mgf of X_1 exist? If so, what is it?

(b) Suppose that $Y \sim \mathrm{NB}(r, s)$; that is,

$$P(Y = y) = \binom{r+y-1}{y} s^r\,(1-s)^y \;\text{ for }\; y = 0, 1, 2, \dots$$

where $s \in (0, 1)$ and $r \in \{1, 2, 3, \dots\}$. Find the mgf of Y.

(c) Find the pmf of the random variable $Z = \sum_{i=1}^n X_i$.

(d) Find the ML estimator of $g(p) = p(1-p)$, call it $\widehat{g(p)}$.

(e) Is $\widehat{g(p)}$ the best unbiased estimator of $g(p)$? If not, find the best unbiased estimator of $g(p)$.

5. Let $X_1, \ldots, X_n$ be iid Uniform$(\theta, \theta+1)$ where $\theta \geq 0$.

(a) Show that $(X_{(1)}, X_{(n)})$ is a sufficient statistic. Is it complete?

(b) Find the cdf of X_1, the cdf of $X_{(1)}$ and the joint pdf of $X_{(1)}$ and $X_{(n)}$.

For the remainder of this question, we consider testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$ using the test with rejection region

$$R = \left\{ x_1, \ldots, x_n : x_{(1)} > k \text{ or } x_{(n)} > 1 \right\} .$$

(c) Find $k \in (0, 1)$ such that this test is level α where $\alpha \in (0, 1)$.

(d) Show that the power of this test is 1 when $\theta \geq k$.

(e) Find the power function of the test. Hint #1: If $\theta < k$, then it must be the case that $0 < \theta < k < 1 < \theta + 1$. Hint #2: $A \cup B = A \cup (\overline{A}$

7. Consider a situation where we have $n = 24$ observations on a response Y and two predictors, X_1 and X_2. Here are three different models along with the actual sum of squares error for each:

$$\begin{array}{lll} \text{Model 1:} & \mathrm{E}Y = \alpha_0 + \alpha_1 X_1 & SSE_1 = 460 \\ \text{Model 2:} & \mathrm{E}Y = \beta_0 + \beta_1 X_2 & SSE_2 = 696 \\ \text{Model 3:} & \mathrm{E}Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 & SSE_3 = 459 \end{array}$$

You may use the fact that $\sum_i (Y_i - \overline{Y}.)^2 = 41257$.

(a) Test whether X_2 is related to Y, ignoring X_1. Write out the null and alternative hypotheses, rejection region, and conclusion, based on the $\alpha = 0.05$ significance level.

(b) Test whether X_2 is related to Y, after controlling for X_1. Write out the null and alternative hypotheses, rejection region, and conclusion, based on the $\alpha = 0.05$ significance level.

(c) Give the coefficient of determination between Y and X_2.

(d) Give the coefficient of partial determination between Y and X_2 after controlling for X_1. (Recall that the coefficient of partial determination between Y and a predictor variable is the fraction of the variation in Y that is not explained by the remaining k-1 predictors that is explained by the the current predictor when it is added to the model containing the other k-1 predictors.)

8. An experiment is conducted with 4 replicates at each of 5 levels of an independent variable $(0, 5, 10, 15, 20)$. The experimenter is interested in two models. Model 1 presumes a linear relation between the mean response and the independent variable. Model 2 allows for the mean response to differ among the levels of the independent variable, but does not presume a shape to the relationship.

$$\text{Model 1: } \mathrm{E}Y_{ij} = \beta_0 + \beta_1 X_i \qquad \text{Model 2: } \mathrm{E}Y_{ij} = \mu_i$$

(a) Derive the least squares normal equations under Model 1.

(b) Derive the least squares estimates of the group means under Model 2.

(c) The group means (variances) are: 25 (2), 30 (4), 20 (3), 35 (4), 40 (3). Give the Analysis of Variance under Model 2.

(d) Give the least squares estimates of β_0 and β_1 under Model 1.

9. An experimenter wishes to compare four types of ink in terms of fading. She selects 16 sheets of paper at random from a large ream of paper, and marks each sheet with each type of ink, and measures the amount of fading on each mark.

(a) Treating this as a randomized complete block design, where the ink types are fixed effects and the sheets of paper are random effects, write out the model, assuming sheet effects are independent and normally distributed and that random errors are independent and normally distributed. Further, assume that the random errors are independent of sheet effects. Note that Y_{ij} is the measurement made when ink i is applied to sheet j.

(b) Give the covariance structure of the data.

(c) A partial ANOVA table is given below. Test whether the ink means differ at the $\alpha = 0.05$ significance level. Write out the null and alternative hypotheses, rejection region, and conclusion.

Source	df	SS
Ink	3	1800
Sheet		30000
Error		
Total		36300

(d) Give the minimum significant difference based on Bonferroni's approach to determine whether two ink types differ in terms of fading with an experimentwise error rate of $\alpha = 0.05$ significance level.

10. Consider the matrix form of the linear regression model: $Y = X\beta + \varepsilon$ where $\varepsilon \sim \mathrm{N}(0, \sigma^2 I)$. The least squares estimator of β is, of course, $\hat{\beta} = (X'X)^{-1}X'Y$.

(a) Derive the mean vector and variance-covariance matrix of $\hat{\beta}$.

(b) Write out the vector of predicted values $\hat{Y}$ as a function of $\hat{\beta}$. Derive its mean vector and variance-covariance matrix.

(c) Write out the vector of residuals, e. Derive its mean vector and variance-covariance matrix.

(d) Show that the vector of residuals and the vector of predicted values are orthogonal; that is, that the sum of the products of the fitted values and residuals is 0.

(e) Was the restriction on the residual vector necessary in parts (a)-(d)? **Yes** or **No**